Number 14 February 2016

Centre for Development Impact
PRACTICE PAPER
Innovation and learning in impact evaluation

Balancing Inclusiveness, Rigour and Feasibility: Insights from Participatory Impact Evaluations in Ghana and Vietnam

Abstract This paper by Adinda Van Hemelrijck and Irene Guijt explores how impact evaluation can live up to standards broader than statistical rigour in ways that address challenges of complexity and enable stakeholders to engage meaningfully. A Participatory Impact Assessment and Learning Approach (PIALA) was piloted to assess and debate the impacts on rural poverty of two government programmes in Vietnam and Ghana funded by the International Fund for Agricultural Development (IFAD). We discuss the trade-offs between *rigour*, *inclusiveness* and *feasibility* encountered in these two pilots. Trade-offs occur in every impact evaluation aiming for more than reductionist rigour, but the pilots suggest that they can be reduced by building sufficient research and learning capacity.

1 Introduction

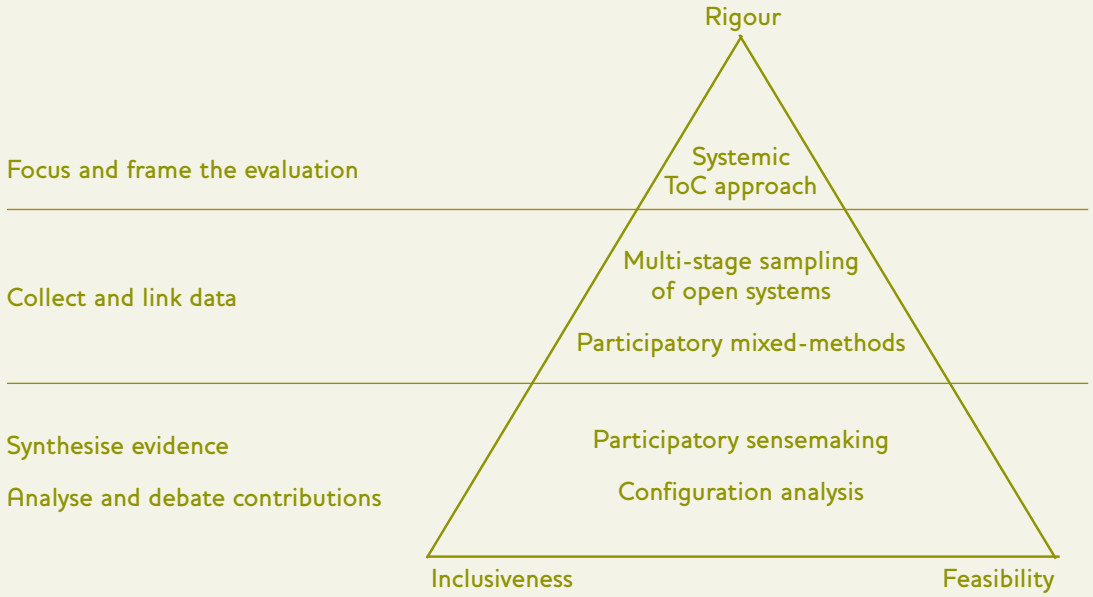
The past ten years have seen a surge in interest and investment in impact evaluation in development. Bulletproof numbers are required to justify programme investments at scale. Credible explanations of observed changes are needed to influence national policy and local responsibility for greater impact. Large programmes with big investments, though, are increasingly complex and political (Wild *et al.* 2015). Interventions are less standardised, stakeholders are more diverse, influences are more dense, problems are more intertwined and systemic, solutions are less straightforward, and changes are emergent and less predictable (Burns 2014a; Woolcock 2013). Additionally, Sustainable Development Goals (SDGs) are adding demands for greater inclusiveness and sustainability as well as effectiveness, forcing a rethink of impact evaluation.

Mainstream practice is still dominated by the need for numbers that attribute impact to investments. Counterfactual-based approaches using statistical and (quasi-)experimental methods fit this need (White 2009). Generally, these are very costly and difficult to pursue in complex environments; they do not explain impacts or enable stakeholders to engage and learn. Rigour is purely statistical and method-specific, thus inadequate for other (including participatory) methods (Stern *et al.* 2012). *How then to ensure that impact evaluation of complex programmes is rigorous and inclusive, and triggers reflection and learning about contributions, while also remaining feasible?*

Responding to this challenge, the International Fund for Agricultural Development (IFAD)[1] and the Gates Foundation (BMGF) commissioned the development of a Participatory Impact Assessment and Learning Approach (PIALA) that could help IFAD and partners rigorously and collaboratively *assess*, *explain* and *debate* their contributions to reducing rural poverty (IFAD and BMGF 2013b).[2] PIALA was designed and piloted around standards of *rigour*, *inclusiveness* and *feasibility* that are considered vital to create impact evaluations of value.

Rigour[3] in this approach refers to the quality of thought put into the methodological design and conduct of every step in the evaluation – including sampling, triangulation

Figure 1 **PIALA design elements**



Source: Presentation given by the authors at the IDEAS Global Assembly, 29 October 2015, Bangkok.

of methods, facilitation of processes, data collation, cross-validation and causal analysis. This is to ensure consistency and responsiveness to the purposes and constraints of the evaluation, necessary to establish sufficient confidence among stakeholders in its findings and conclusions (Rogers 2009; Stern *et al.* 2012). *Inclusiveness* involves meaningful engagement of stakeholders with diverse perspectives, which has an intrinsic empowering value while also enhancing credibility of the evaluation through triangulation and cross-validation of evidence. *Feasibility* concerns the budget and capacity needed to meet expectations of rigour and inclusiveness and to enhance learning (Chambers 2015).

PIALA was first piloted in Vietnam (IFAD and BMGF 2014) and then in Ghana (MOFA/GOG, IFAD and BMGF 2015). Both pilots used these three standards for framing the evaluation: data collection and linking; synthesising findings; and analysing and debating programme contributions. Insights from the first pilot enabled the second to address some of the challenges encountered. This paper describes key trade-offs and lessons. First, we present PIALA as a response to the main challenges of impact evaluation in complex environments. Then, we discuss our insights about the possible trade-offs. We conclude by presenting reflections on how to balance rigour, inclusiveness and feasibility.

2 PIALA's response to challenges in impact evaluation

Challenges in impact evaluation

IFAD-funded government programmes are implemented in rather complex ways and environments that challenge mainstream evaluation. Four challenges are common to many contexts.

First, clean 'control' groups are rarely found and quantitative estimation of net attributable impacts on 'treated' compared to 'control' groups is often impossible or inadequate. This is the case in IFAD programmes where institutional and policy work 'contaminate' entire populations, where other donors and influences augment 'causal density', where self-targeting mechanisms make 'treatments' highly diverse, and where innovations have emergent results (Woolcock 2013). Hence the need for different ways of arriving at rigorous causal inference that better fit complex environments (Befani 2012; Guijt and Roche 2014).

Second, programme managers and even funders can feel threatened by traditional types of evaluation that focus on performance against pre-set targets, whereas in complex environments there is less control over results (e.g. similar processes can lead to different outcomes). This hinders solid debate and learning about impact. By seeking to understand programme contributions to impact, alongside many other influences, and by taking a broader systemic perspective, fear of failure can partially be sidestepped (Eyben *et al.* 2015).

Third, it is important to analyse and understand development impacts more systemically when seeking transformational or systemic change that is more inclusive and sustainable and grounded in rights and democracy (Eyben 2008). The tendency in mainstream evaluation practice is to slice programmes into measurable parts and then look at intervention and effect for each in isolation (Befani, Ramalingam and Stern 2015). In Ghana, for instance, in some studies of specific programme mechanisms conducted before the PIALA study, such a reductionist perspective resulted in quite perilous

Table 1 **PIALA methods and processes**

Methods and processes	Purposes
Focusing and framing the evaluation	
1 Outlining of **design options** and budget implications (*full scale–full scope; full scale–limited scope;* or *limited scale–full scope*)	■ Enable commissioners to decide on scope and scale of evaluation
2 Reconstruction and visualisation of programme **Theory of Change (ToC)**	■ Identify causal claims and assumptions ■ Formulate evaluation questions ■ Create shared understanding among stakeholders of programme theory and broader influences
Collecting and linking data	
3 Multi-stage sampling with 'open systems' as principle sample unit (e.g. value chain systems) **4** Selection of methods for data collection, and drafting of 'how-to' guidance and templates for each method and for quality monitoring	■ Enable systemic inquiry and comparative analysis of impact on livelihoods and household poverty ■ Enable rigorous use of methods and facilitation of processes ■ Enable systematic data quality monitoring and reflective practice
5 Data collection on changes and causes in household food and income through: ■ **Household survey** ■ **Generic change analysis** in gender-specific groups (*social mapping, timeline, wealth and wellbeing ranking, causal flow mapping*)	■ Collect and triangulate data on impacts ■ With intended beneficiaries visually reconstruct and discuss causal flow of changes in livelihoods affecting household wealth and wellbeing
6 Data collection on livelihood changes and causes through: ■ **Generic change analysis** (see above) ■ **Livelihood analysis** in gender-specific groups (*livelihood change matrix, causal flow mapping, SenseMaker*)	■ Collect and triangulate data on effects of livelihood changes on household food and income ■ Visualise and discuss with intended beneficiaries causal flow of changes and causes in different areas affecting their livelihood
7 Data collection on reach and effects of selected programme mechanisms through: ■ **Livelihood analysis** (see above) ■ **Constituent Feedback (CF)** in mixed groups (questionnaire for discussing and anonymous scoring) ■ **Semi-structured interviews** with service providers and officials (*CF-linked questionnaire*)	■ Collect and triangulate data on effects of programme mechanisms on changes and causes in various areas affecting livelihoods ■ With intended beneficiaries discuss and anonymously score reach, benefits, outcomes of mechanisms
8 Data linking and quality monitoring using a standard data collection tool and questionnaire for team reflections on quality of methods, processes and evidence using a standard questionnaire	■ Enable instant data processing and cross-checking to identify gaps and weaknesses ■ Ensure robust evidence (inclusive, sufficient, consistent, rigorous)
Synthesising evidence and analysing and debating programme contributions	
9 Local and national **participatory sensemaking** using a workshop model consisting of design principles and methods for enabling voice and facilitating cross-validation and contribution scoring **10 Configurational analysis** using standardised data collation and scoring tools	■ Probe to fill remaining data gaps ■ Enable stakeholders to understand impact systemically ■ Engage stakeholders in valuing programme contributions and identifying priority investment areas

Source: Drawn from the Root and Tuber Improvement and Marketing Programme (RTIMP) impact evaluation report in the Ghana pilot (cf. MOFA/GOG, IFAD and BMGF 2015).

recommendations – e.g. scaling up production in the absence of markets. A systemic approach therefore must inquire not just 'what works' but also 'why' and 'how' in order to understand likely sustainability (Burns 2014a).

Fourth, greater inclusiveness and sustainability of development outcomes and impacts requires stakeholders to learn systemically and adapt responsibly in complex environments, thus understanding broader influences and individual and collective responsibilities (Bawden 2010). Mohan and Hickey (2004) plead for critical participation based on 'rigorous, debated and contested' evidence from participatory research. Mainstream evaluation largely fails to meaningfully engage stakeholders (particularly beneficiaries), assuming that asking them directly about attribution is highly susceptible to bias (Copestake 2013). Mixing different methods, including participatory methods, to investigate contribution (broader than attribution) is suggested as a better way forward (Bamberger 2012; Stern *et al.* 2012).

PIALA design elements

To address these challenges, PIALA is constructed around five key design elements (presented in Figure 1): a systemic theory of change (ToC); multi-stage sampling centred on 'open systems'; participatory mixed-methods; configurational analysis; and participatory sensemaking. Table 1 shows how these design elements were translated into a practical sequence of methods.

The ToC approach is most critical, as it provides the structure for the entire evaluation. This involves visualising a programme's ToC showing the systemic links and feedback loops between different programme components and mechanisms and other influences, and their collective outcomes and impacts. Reconstruction and visualisation happens through iterative discussions with stakeholders. Data collection methods are chosen related to the ToC to inquire about cascading causes and effects (expected and unexpected, positive and negative), from household impacts down to programme mechanisms. This process permits rigorous and systematic assessment of multiple interacting causal claims together with stakeholders. In sensemaking workshops, participants construct a causal flow diagram with the evidence that mirrors – and thus validates or refutes – the ToC. This enables stakeholders to collectively value programme contributions along two scales from '*strong contribution*' to '*contribution overrun by other influences*', and from '*positive*' to '*negative*' impact (IFAD and BMGF 2015; Van Hemelrijck 2013).

About the pilot cases

PIALA was piloted in the Developing Business with the Rural Poor (DBRP) Programme in Vietnam and the Root and Tuber Improvement and Marketing Programme (RTIMP) in Ghana. Both programmes focused on improving livelihoods and increasing incomes as part of sustainable and equitable poverty reduction through enhancing smallholders' capacity to commercialise, and linking local businesses to markets and industries. While the DBRP focused on diversified short value chains, the RTIMP sought to develop longer commodity chains linked to national and export markets and industries. Both programmes were at completion.

- The DBRP was implemented from 2008 to 2014 in two provinces (Cao Bằng and Bến Tre) with a budget of US$51m, including US$36m from IFAD. The evaluation was conducted in 2013 at a cost of US$90,000. It covered five years of work in Bến Tre province only, where the programme was implemented in 50 of 164 communes in eight of nine districts.

- The RTIMP was implemented from 2007 to 2015 in 106 of Ghana's 216 districts across all ten provinces, with a total budget of US$24m, of which US$19m was from IFAD. The evaluation took place after programme completion in 2015 and was conducted countrywide at a cost of US$233,000. It covered the period from 2010, after the mid-term review.

3 Reflections about trade-offs in practice[4]

Focusing and framing the evaluation

In the first phase of an impact evaluation, clarity and shared understanding must be created among key stakeholders about what is to be evaluated and for what purposes. This involves considering different design options, identifying causal claims and assumptions, deciding on focus and questions, and specifying criteria and standards for evaluation (BetterEvaluation 2014). Using a systemic ToC approach makes it possible to do this rigorously and collaboratively with stakeholders, yet requires sufficient time and budget.

Ensuring ownership of the ToC (rigour and inclusiveness versus feasibility)

Inclusiveness is critical to ensure that the ToC is rigorously used for framing and focusing the evaluation, creating ownership, and facilitating inclusiveness in the analysis stage too. Yet this also makes feasibility more elusive.

In Vietnam, due to budget and time limitations, we made three choices that compromised inclusiveness and subsequent rigorous use of the ToC. First, we organised a brief workshop with the programme steering committee and managers only to discuss the programme logic and expectations for evaluation prior to reconstructing the ToC. Other stakeholders were not involved. Second, we did not allow sufficient time for the researchers to engage in desk review and national stakeholder interviews to inform the ToC process. Instead, we prioritised their involvement in refining field methodology. Third, we assumed that the broader evaluation questions outlined in the PIALA strategy paper (IFAD and BMGF 2013b) would be sufficient

to guide fieldwork and to focus analysis. However, the evaluation scope remained too wide, and researchers and stakeholders showed limited ownership of the ToC, making it difficult for them to relate the evidence to the causal narrative. The resulting lack of shared understanding of evaluation frame and focus hindered arriving at greater precision of evaluation findings and agreement on strategic directions for the next programme.

Armed with these insights, we required the researchers in Ghana to conduct the desk review and lead on the ToC reconstruction and visualisation. This helped them thoroughly understand the programme's impact and contribution claims. A design workshop convened stakeholders to build shared understanding of the ToC and agree on the mechanisms, assumptions and questions the evaluation should focus on. This more robust and collaborative process laid the foundation for the entire evaluation.

Value for money of design options (rigour and inclusiveness versus feasibility)

Purpose and budget strongly influence design options. The more rigorous an evaluation design, the less feasible it is for low-capacity, low-resource situations.

In Ghana, prior to signing contracts, commissioners were offered three design options explaining what value for money they could expect from each: 'full scope – full scale', 'limited scope – full scale', or 'full scope – limited scale'. Scale refers to the size of the principal sample of 'systems' (in the PIALA pilots: value chain systems) from which households for survey and intended beneficiaries for participatory research are subsampled. Scope refers to coverage of programme components and mechanisms reflected in the ToC. In Ghana, commissioners chose the first and most expensive option for reporting and learning purposes (MOFA/GOG, IFAD and BMGF 2015). When operating on a shoestring budget, either scope or scale may need to be downsized.

A 'full scope – limited scale' design emphasises learning about the programme's contribution to impact in select cases under specific conditions. Fieldwork and analysis are less resource-intensive, but findings are not generalisable (unless the programme itself is case-based). In a 'limited scope – full scale' design, the purpose is to learn about the effects of selective programme mechanisms on the entire population. A ToC approach is not mandatory, saving time and money; but this runs the risk of arriving at wrong conclusions and limiting stakeholders' learning. For example, a cost-effectiveness study of Farmer Field Forums in Ghana recommended scaling up because of the high adoption of new technologies (MOFA 2014), while the PIALA study showed that in Ghana's downward conjuncture, this contributed to market saturation, which negatively affected livelihoods.

In Vietnam, this trade-off was not yet well understood. We believed that participatory research in a subsample of the already *small-scale* sample of villages where a representative household survey was conducted would be sufficient to conduct a full scope inquiry of programme contribution to impact for the entire province. Rigorous causal inference was hindered by our limited confidence in linking *not case-specific* household-level findings to *case-specific* participatory research findings, and our inability to generalise participatory research findings regarding the effects of programme mechanisms on livelihoods.

Collecting and linking data

In this phase, decisions are made about how to obtain and collate data. This includes determining the sample frame and selecting the mix of methods that must ensure data sufficiency, data linking and quality monitoring while also enabling meaningful participation (BetterEvaluation 2014). The pilots experienced multiple trade-offs at this stage. Those discussed here are: (1) sampling of open systems versus well-defined units; (2) prioritising depth or breadth of inquiry; and (3) independent field mobilisation versus involving programme staff.

Sampling of open systems versus well-defined units (rigour versus feasibility)

A systemic perspective for analysing contributions to impact requires identifying the system that forms the main unit for the sample frame. In complex development contexts, systems have open boundaries, meaning they interact with their environment (Burns 2014b; Humphrey 2014). Open systems are difficult to discern and sample. Time and budget constraints – particularly if there is no shared understanding of the system – may lead to compromising on rigour and finding proxy sample units.

In Vietnam, two factors made it ostensibly easy to determine the sampling frame. First, the programme focused on developing short value chains close to the farmers and involving just a few local actors. Hence we assumed that the villages formed an adequate proxy as main sample unit. Second, demographic data from which to sample households were readily available from provincial government agencies. However, because we had not clearly identified the value chains and subsampled the households from these, it was difficult to probe for systemic interactions and link data on changes in business environment and local capacity of service providers to changes in livelihoods within these value chains, thus compromising on analytical rigour.

Learning from this, much more work was put into the sampling frame in Ghana. The RTIMP developed long commodity chains, consisting of many supply chains around the country. The evaluation focused on four commodities, so had four different populations of supply chains from which to sample. Supply chains are loose geographic areas where smallholders supply raw products to a small

enterprise or industrial off-taker manufacturing higher-value products for bigger markets. The chains interact and overlap geographically and administratively. Hence they were hard to discern and, in practice, often differed from what was sampled on paper. Ensuring that evidence collected on these entities remained comparable entailed much creativity and coordination.

Also, there were no population lists available for sampling households. Census lists and national data could not be matched with the sampled supply chain areas. Budget limits did not permit extra fieldwork to construct the lists. Instead, we systematically sampled every fifth or tenth household in a straight or zigzag line from a central point of the main community in each supply chain area. Separately, focus group participants were sampled from beneficiary lists (where available) or from lists created by applying a snowballing technique.

Independent field mobilisation versus programme staff engagement (rigour versus feasibility)

Field mobilisation is best undertaken independently from the programme to avoid positive bias. If this is impossible, strong facilitation skills are needed to minimise undue influence by programme staff.

In Vietnam, the evaluation took place in one province and communities were easily accessible. Researchers had manageable travel distances. Focus group discussions were easy to organise within the villages, so participants did not have to travel. Local transportation and mobilisation was organised by local officials and programme staff, which made it more difficult to maintain independence and avoid interference. On the other hand, staff and officials were more engaged as the evaluation unfolded.

In Ghana, the researchers took responsibility for transportation and mobilisation to warrant greater independence. Staff and officials were only engaged in interviews and workshops. However, budget and schedule were challenged by the scale of the evaluation (national), very poor infrastructure, remoteness of communities, and spread of communities in the supply chain areas. Researchers and participants had to travel far on poor roads to conduct the focus group discussions. Finding central locations suitable for these meetings that were known and trusted by all participants from all communities proved cumbersome.

Depth versus breadth of inquiry (inclusiveness versus rigour and feasibility)

PIALA's mixed-methods approach pursues depth, through focused participatory inquiry of 'open systems', and breadth, through representative household surveys. To enable data linking, households are sampled within the sample of these 'systems'. However, there are trade-offs when using this approach at scale, as participatory research becomes onerous and limited resources can further compromise rigour.

In Vietnam, the researchers were comfortable with survey statistics, but under pressure of time they struggled with linking and triangulating data from participatory research. Their quantitative background and limited understanding of the ToC, with methods focusing more on generic changes, meant that qualitative data capture and collation was not as structured and rigorous as in Ghana. The large amount of qualitative data was overwhelming and left them with little time during fieldwork for daily reflections on data quality and interim findings. Methods and tools for data linking and quality monitoring were insufficient to adequately guide them. We also underestimated post-collection tasks. Consequently, data collation tables and village sensemaking workshop reports were submitted long after fieldwork ended, making verification and data gap-filling impossible – in turn challenging aggregated analysis and causal inference.

In Ghana, the evaluation did not sacrifice depth or breadth; all questions were answered adequately. The researchers were selected based on their experience of mixed-methods and participatory research. With methods tightly focused on causal links in the ToC, their solid grasp of the ToC helped them facilitate the processes and collate and triangulate data more rigorously. Data collation and quality monitoring was undertaken daily and systematically. Data collation was structured according to the causal claims and links in the ToC. Quality monitoring focused on assessing the inclusiveness of participatory processes and the robustness of emerging evidence for each of the causal links. Teams were able to identify data gaps and weaknesses in good time, and prepare well for the district sensemaking workshops.

Yet the fatigue of six weeks' fieldwork undoubtedly affected their ability to produce persistently high quality. No real answer exists to this: more teams in parallel would have led to greater variation, while extending fieldwork to allow more breaks would have affected researchers' levels of commitment. The trade-off between breadth and depth is clearly less prominent, however, when working with highly competent and motivated research team leaders and a research coordinator who takes pride in high-quality research. In contexts where local research capacity is weaker, feasibility of a 'full scope – full scale' design becomes questionable. Options include more investment in training, coaching and supervision and/or less ambitious designs.

Synthesising evidence, and analysing and debating contribution claims

In this phase, data are 'zipped up' again along the ToC to show what evidence upholds or refutes the assumed contribution claims. This leads to answering causal questions about what has produced which observed outcomes and impacts, for whom and why (BetterEvaluation 2014). Participatory sensemaking and configurational analysis (cf. Table 1) form the backbone of this phase.

Figure 2 **Part of the RTIMP configurational analysis**

					Contribution Claim of RTIMP Component 3 → Enhanced Processing (O3)			Contribution Claim of RTIMP Component 2 → Enhanced Production (O2)			Contribution Claim of RTIMP Component 1 → Enhanced Market-Linking (O3)			Contributions of RTIMP Components 1, 2 and 3 → Improved Livelihoods (I2)	
	DSF	FFF	GPC	MEF	MEF (M3c)+C1a +M3b → C3c	GPC (M3b)+C3c → C3b → O3	Evidence Strength	FFF M2a+M2b+ (M2c) → C2a	C2a+C2b → O2	Evidence Strength	M1c+M1b+ O3+O2 +O1 → C1b	DSF C1a+(M1) → O1	Evidence Strength	O1+O2+ O3 → I2	Evidence Strength
Tano North (Apesika) (CZ)	1	1	1	1	3	6	5	5	5	5	4	4	5	5	5
Techiman (CZ)	1	1	1	1	4	5	5	5	5	5	4	4	5	5	5
Gomoa East (SZ)	1	1	1	0	2	5	3	5	5	5	4	4	5	5	6
Assin South (SZ)	1	1	1	1	3	4	4	6	5	4	3	3	4	4	4
Birim Central (CZ)	1	1	1	1	3	3	4	5	5	4	3	4	4	4	5
Nkwanta South (NZ)	1	1	1	0	3	4	5	5	4	5	3	3	5	4	5
Upper West Akim (CZ)	1	1	1	1	2	4	4	5	5	4	3	3	5	4	5
Ashanti Mampong (CZ)	1	1	1	1	3	4	5	5	5	5	3	3	5	4	5
West Gonja (Damongo) (NZ)	1	1	1	0	3	4	5	5	4	5	3	3	5	4	5
Abura Asebu Kwamankese (SZ)	1	1	1	1	3	3	5	5	5	6	3	3	5	4	4
Nanumba North (NZ)	1	1	N/A		N/A			5	5	5	3	3	5	4	5
East Gonja (NZ)	1	1	N/A		N/A			4	3	5	3	3	5	4	5
Central Gonja (NZ)	1	1	N/A		2	3	5	5	4	5	2	2	5	4	5

NZ=Northern Zone
CZ=Central Zone
SZ=Southern Zone

Gari | HQCF | Yam | PCF | Other

Source: MOFA/GOG, IFAD and BMGF (2015).

Classical counterfactual or configurational counterfactual (rigour versus feasibility)

Mainstream impact evaluation assumes that comparative data analysis from treated and non-treated sites is both accessible (thus feasible) and necessary (thus rigorous) to reach generalisable conclusions about impact on rural household poverty. However, where this is not the case (which is quite common), other forms of rigorous analysis are needed.

In the Vietnam pilot, concerns about heterogeneity in programme treatment and sample limitations made it difficult to align findings with the ToC. These concerns and limitations made us hesitant to generalise certain findings regarding programme contribution. Rigour and feasibility appeared as an 'either/or' type of trade-off.

In Ghana, this trade-off was solved by choosing a different causal approach combining 'configurational' with 'generative' perspectives (Punton and Welle 2015; Stern *et al.* 2012). We used systemic heterogeneity as the basis for identifying and analysing programme contributions. Instead of a classic counterfactual inquiry of household-

level impact, we employed a counterfactual approach that looked at the effects of different patterns of treatment that combined presence/non-presence, functional conditions and differentiated effects of programme mechanisms on livelihoods and households.

We developed a configurational analysis method to compare the evidence collected for each causal link in each of the three contribution claims in the ToC across our sample of supply chains (first column in Figure 2). For each supply chain, the formal presence of programme mechanisms such as Farmer Field Forum (FFF) or Micro-Enterprise Fund (MEF) was inputted as a binary code (next four columns in Figure 2). We scored the causal link between the contribution claims and the impact claim (columns on 'livelihood improvements'), and the evidence for this link, on 'strength' and 'consistency'.

Similarly, we scored the evidence for each causal link in which a mechanism operated, and the reach and performance of the mechanism in each contribution claim (columns on 'enhanced processing', 'enhanced production' and 'enhanced market-linking'). The scoring was done based on detailed explanatory evidence collected for each of the claims independently (with different sets of methods and different groups). The analysis then looked at similarities and differences of various configurations of clusters of scores across the supply chains supported by the evidence (MOFA/GOG, IFAD and BMGF 2015).

Not all contexts will allow for such a thorough and detailed configurational analysis, as it requires high-quality data and analytical capacity. While there will always be a tension between rigour and feasibility, the approach can be adapted to bring rigorous analysis within reach, in contexts where classical counterfactual approaches cannot be applied.

Involving international experts or investing in local capacity (rigour versus feasibility)

Undertaking a rigorous aggregated multi-causal analysis demands high-level analytical skills. In contexts where local research institutions do not yet have these skills, conducting impact evaluations of complex programmes such as those funded by IFAD becomes less feasible.

In both pilots, researchers were not experienced with rigorous aggregated multi-causal analysis. Because of the methodological innovation, the authors took responsibility for the final analytical product. Ideally, however, the national research coordinator should undertake the aggregated analysis and final reporting as part of delivery of the evaluation.

To ensure sufficient analytical and reporting capacity, we see two options. The first is to work – as we did in Vietnam – with international impact evaluation specialists leading on final analysis and reporting. This option is not optimal for fostering in-country capacity and responsibility for conducting rigorous impact evaluations.

A second option involves investing in research partnerships with in-country research firms, thus strengthening local competencies. Integrating PIALA in programme design as part of an impact-oriented monitoring and evaluation (M&E) process would increase cost-effectiveness, lay the foundation for better knowledge for policy and decision-making, and create more democratic space for stakeholders to influence decisions (Guijt 2014; Peersman *et al.* forthcoming).

We shifted towards the second option in Ghana, ensuring that those involved in the tendering process fully understood the requirements and including far more days for in-country supportive supervision of the pilot. The research coordinator was strongly involved in the entire evaluation process, leading to stronger ownership and responsibility than in the case of Vietnam. IFAD's country programme manager also actively engaged in design and sensemaking. The Ghana initiative was thus experienced as more of a joint learning journey – a partnership rather than a technical consultancy.

Degree of participation in sensemaking (inclusiveness versus rigour and feasibility)

Rigorous facilitation of participatory sensemaking – i.e. being responsive to local conditions and dynamics, while consistently employing the same set of models and tools in every locality – is essential for enhancing credibility and confidence in evaluation findings (i.e. rigour) as well as generating solid debate and systemic learning among key stakeholders (i.e. inclusiveness). Again, this makes feasibility more elusive.

Engaging beneficiaries, service providers and decision-makers in collective sensemaking of emerging evidence before turning to final analysis and reporting has both instrumental and empowering value (MOFA/GOG, IFAD and BMGF 2015). Doing this in all researched localities and at programme level helps to improve and strengthen the evidence, overcome bias, and create ownership of evaluation findings among stakeholders. For this to succeed, it is crucial to design and facilitate the processes carefully, in ways that enable all participants to critically engage and express their views in the presence of power-holders, and adopt a systemic perspective in valuing programme contributions to impact (cf. Section 2). A participatory sensemaking workshop model was developed that was first piloted in Vietnam and further expanded and improved on in Ghana.

Using this model, in Vietnam, we organised six village-level workshops with 180 participants and one provincial workshop with 100 participants, while in Ghana, there were 23 district workshops with 650 participants and one national workshop with 100 participants. Participants were purposively sampled from the research participants. Beneficiaries comprised more than 70 per cent of those

attending local workshops, and more than 30 per cent at provincial/national workshops. The workshops were quite successful in both pilots. Participants gained a more complete picture of the development processes. There were lively debates about programme contribution to impact and priority areas for future investment. Critical to this success were the time and resources invested in organising the workshops, and the capacity to rigorously design and facilitate them. When operating on a shoestring, the number of workshops and participants may need to be limited. But this undoubtedly has implications for rigour and inclusiveness.

4 Making multiple standards work in impact evaluation

Rigour, *inclusiveness* and *feasibility* are crucial for rethinking impact evaluations to meet the challenges of complexity and sustainability (described in Section 2) and to enable stakeholders to learn and adapt responsibly. But these three standards are not easily compatible. Decisions are influenced by politics, capacities, and contextual factors; trade-offs seem inevitable. In this paper, we have merely given a flavour of the range of possible trade-offs. We encountered many others.

So, if impact evaluation needs to move beyond being driven by reductionist rigour, how does one deal with multiple standards? We conclude with three recommendations.

First, being clear about how to meet each of the standards is crucial. Rigour is arguably the standard around which most was achieved in Ghana as we learned from what had happened in Vietnam and became much clearer about what it entailed. Rigour involved being thorough and careful methodologically and analytically, as well as being thoughtful about whose voices informed the findings. Inclusiveness was considered instrumental to arrive at greater rigour, while also being essential for learning and for influencing future policy and practice. In both PIALA pilots, considerable space was created for stakeholders to cross-validate and debate emerging evidence. Rigour also involved sampling thoroughness and appropriate method selection, which affects the ability to conduct a 'full scale – full scope' evaluation in a way that permits rigorous configurational analysis. This was achieved in Ghana, we argue, despite the classic counterfactual (using control groups) not being feasible.

Second, being clear with commissioners upfront about which standards are essential to serve which purposes helps decisions to reduce problematic trade-offs and accept those that remain inevitable. We sought to pursue all three standards equally as part of the piloting process, but made clear choices. For example, the importance of inclusiveness for both analytical quality and uptake of findings in Ghana led us to invest more in participatory sensemaking, instead of additional participatory data collection on poverty characteristics for designing the household survey. Another example from Vietnam involved the unavoidable presence of programme staff during fieldwork, which was necessary to make it feasible. Undoubtedly, this must have had some influence on what villagers chose to share. Yet we were able to mitigate excesses by rigorous facilitation and triangulation of different processes. These examples show the win-wins that can be achieved from putting serious thought into balancing *rigour*, *inclusiveness* and *feasibility* by carefully reflecting on potential losses in *value for money* if one were to be prioritised over the other.

Third, anticipating possible trade-offs while searching for win-wins is worthwhile to help think of the critical competencies needed to conduct the evaluation. Having learned much in Vietnam, we discussed in detail where and how we needed to fare considerably better in Ghana. Competencies to handle participatory processes and mixed methods proved critical. In Vietnam, researchers struggled with integrating the large amounts of qualitative and quantitative data, and were also challenged by the more open-ended design. In Ghana, researchers were well versed in research involving participatory processes and large amounts of qualitative and quantitative evidence, and were not restrained by the confirmative design to probe for alternative explanations. Having a statistician on the team gave sufficient confidence.

To conclude, trade-offs are clearly not absolute, and it is worthwhile exploring win-wins in any context to reduce losses and enhance the evaluation's *value for money*. One does not have to forfeit inclusiveness completely if rigour is deemed a non-negotiable. Nor does rigour have to be compromised totally when budgets and capacities are restricted. Each of the standards can be conceived of as a gradient. For example, inclusiveness can be approached from a minimalist perspective – ensuring enough to cross-validate key findings but perhaps cutting short on the aspiration of more collective learning and empowering forms of inclusiveness. Similarly, operating constraints may mean that it is not feasible to pursue more detailed surveys in larger samples to build more airtight statistical rigour, yet still permit building sufficient confidence in findings to stand up to scrutiny.

Being inclusive and rigorous does, however, make PIALA more demanding of time, capacity and budget. Evaluations of smaller programmes with smaller budgets and limited scale will not need large samples and are therefore likely to be cheaper, but may still produce little value without sufficient capacity. Arguably, impact evaluation is always difficult, so there are no real shortcuts. But win-wins are more likely where quality standards are clearly defined and where sufficient guidance is provided for every step in the evaluation process – and above all, where there is a commitment to building learning and research capacity.

Notes

1 IFAD is a specialist UN agency providing loans and support to governments for smallholder agricultural development.
2 The piloting of PIALA was made possible with financing from IFAD and BMGF. However, this paper does not represent the views of the funders; it presents views of the individual authors.
3 In dominant evaluation practice, rigour connotes the controlled avoidance of bias through statistical procedure (Befani, Barnett and Stern 2014). This narrow definition is inadequate for mixed participatory methods evaluations. The premise is that bias cannot be avoided by a single method or procedure but can be mitigated through triangulation of different methods and perspectives (Camfield, Duvendack and Palmer-Jones 2014).
4 The paper builds on methodological reflections conducted with stakeholders and researchers that were documented by one of the authors (Adinda Van Hemelrijck) as part of the IFAD and BMGF innovation project and her doctoral study (cf. IFAD and BMGF 2013a, 2015; Van Hemelrijck 2014).

References

Bamberger, M. (2012) *Introduction to Mixed Methods in Impact Evaluation,* Guidance Note 3, Washington DC: InterAction

Bawden, R. (2010) 'Messy Issues, Worldviews and Systemic Competencies', in C. Blackmore (ed.), *Social Learning Systems and Communities of Practice,* Springer, 89–101

Befani, B. (2012) *Models of Causality and Causal Inference,* review prepared as part of the DFID study, *Broadening the Range of Designs and Methods for Impact Evaluation*, London: Department for International Development

Befani, B.; Barnett, C. and Stern, E. (2014) 'Introduction – Rethinking Impact Evaluation for Development', *IDS Bulletin* 45.6: 1–5

Befani, B.; Ramalingam, B. and Stern, E. (2015) 'Introduction – Towards Systemic Approaches to Evaluation and Impact', *IDS Bulletin* 46.1: 1–6

BetterEvaluation (2014) *BetterEvaluation Rainbow Framework and Planning Tool*, http://betterevaluation.org/resources/download_the_Rainbow_Framework (accessed 11 December 2015)

Burns, D. (2014a) *Assessing Impact in Dynamic and Complex Environments: Systemic Action Research and Participatory Systemic Inquiry,* CDI Practice Paper 8, Brighton: IDS, www.ids.ac.uk/publication/assessing-impact-in-dynamic-and-complex-environments-systemic-action-research-and-participatory-systemic-inquiry (accessed 11 December 2015)

Burns, D. (2014b) 'Systemic Action Research: Changing System Dynamics to Support Sustainable Change', *Action Research* 12.1: 3–18

Camfield, L.; Duvendack, M. and Palmer-Jones, R. (2014) 'Things You Wanted to Know about Bias in Evaluations but Never Dared to Think', *IDS Bulletin* 45.6: 49–64

Chambers, R. (2015) 'Inclusive Rigour for Complexity', *Journal of Development Effectiveness* 7.3: 327–35

Copestake, J. (2013) *Credible Impact Evaluation in Complex Contexts: Confirmatory and Exploratory Approaches*, Bath: Centre for Development Studies, University of Bath

Eyben, R. (2008) *Power, Mutual Accountability and Responsibility in the Practice of International Aid: A Relational Approach,* Working Paper 305, Brighton: IDS

Eyben, R.; Guijt, I.; Roche, C. and Shutt, C. (2015) *The Politics of Evidence and Results in International Development: Playing the Game to Change the Rules?*, Rugby: Practical Action Publishing

Guijt, I. (2014) *Participatory Approaches,* Methodological Briefs, Impact Evaluation 5, Florence: UNICEF Office of Research – Innocenti

Guijt, I. and Roche, C. (2014) 'Does Impact Evaluation in Development Matter? Well, It Depends What It's For!', *European Journal of Development Research* 26.1: 46–54

Humphrey, J. (2014) *Market Systems Approaches: A Literature Review,* Beam Exchange, Brighton: IDS

IFAD and BMGF (2015) *Improved Learning Initiative for the Design of a Participatory Impact Assessment and Learning Approach (PIALA): Methodological Reflections following the Second PIALA Pilot in Ghana,* Rome: IFAD and BMGF

IFAD and BMGF (2014) *Impact Assessment of the 'Doing Business with the Rural Poor' Project in Bến Tre, Vietnam (2008–2013). Report of the Pilot Application of a Participatory Impact Assessment and Learning Approach,* Rome: IFAD and BMGF

IFAD and BMGF (2013a) *Improved Learning Initiative for the Design of a Participatory Impact Assessment and Learning Approach (PIALA): Insights and Lessons Learned from the Reflections on the PIALA Piloting in Vietnam,* Rome: IFAD and BMGF

IFAD and BMGF (2013b) 'PIALA Research Strategy. Improved Learning Initiative', Internal Document, International Fund for Agricultural Development, Rome: IFAD

MOFA (2014) *Impact Assessment of FFF on Farmer Beneficiaries of RTIMP. Final Report,* Accra: Government of Ghana, Ministry of Food and Agriculture

MOFA/GOG, IFAD and BMGF (2015) *Final Report on the Participatory Impact Evaluation of the Root & Tuber Improvement & Marketing Program (RTIMP) conducted by PDA with support from the MOFA/GoG. Pilot Application of a Participatory Impact Assessment and Learning Approach (PIALA) Developed with Support from IFAD and the BMGF*

Mohan, G. and Hickey, S. (2004) 'Relocating Participation within a Radical Politics of Development: Critical Modernism and Citizenship', in S. Hickey and G. Mohan (eds), *Participation: from Tyranny to Transformation? Exploring New Approaches to Participation in Development*, London: Zed Books, 59–74

Peersman, G.; Rogers, P.; Guijt, I.; Hearn, S.; Pasanen, T. and Buffardi, A. *et al.* (forthcoming) *Guidance on When and How to Develop an Impact-Oriented Monitoring and Evaluation System,* London: Overseas Development Institute

Punton, M. and Welle, K. (2015) *Straws-in-the-wind, Hoops and Smoking Guns: What can Process Tracing Offer to Impact Evaluation?*, CDI Practice Paper 10, Brighton: IDS, www.ids.ac.uk/publication/straws-in-the-wind-hoops-and-smoking-guns-what-can-process-tracing-offer-to-impact-evaluation (accessed 11 December 2015)

Rogers, P. (2009) 'Matching Impact Evaluation Design to the Nature of the Intervention and the Purpose of the Evaluation', *Journal of Development Effectiveness* 1.3: 217–26

Stern, E.; Stame, N.; Mayne, J.; Forss, K.; Davies, R. and Befani, B. (2012) *Broadening the Range of Designs and Methods for Impact Evaluations,* Working Paper 38, London: Department for International Development

Van Hemelrijck, A. (2014) 'Understanding "Rigor": Challenges in Impact Evaluation of Transformational Development', PhD Research Outline Paper (Revised), Brighton: IDS

Van Hemelrijck, A. (2013) 'Powerful Beyond Measure? Measuring Complex Systemic Change in Collaborative Settings', in J. Servaes (ed.), *Sustainability, Participation and Culture in Communication: Theory and Praxis,* Bristol: Intellect

White, H. (2009) *Some Reflections on Current Debates in Impact Evaluation,* Working Paper 2009-1, International Initiative for Impact Evaluation (3ie)

Wild, L.; Booth, D.; Cummings, C.; Foresti, M. and Wales, J. (2015) *Adapting Development: Improving Services to the Poor,* London: Overseas Development Institute, www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/9437.pdf (accessed 11 December 2015)

Woolcock, M. (2013) 'Using Case Studies to Explore the External Validity of "Complex" Development Interventions', *Evaluation* 19.3: 229–48

> "… trade-offs clearly are not absolute and win-wins [are] worthwhile exploring… to reduce losses and enhance the evaluation's value-for-money. One does not have to forfeit inclusiveness completely if rigour is deemed a non-negotiable. Nor does rigour have to be compromised totally when budgets and capacities are restricted. Each of the standards can be conceived of as a gradient."

Centre for Development Impact (CDI)

The Centre is a collaboration between IDS (www.ids.ac.uk), Itad (www.itad.com) and the University of East Anglia (www.uea.ac.uk).

The Centre aims to contribute to innovation and excellence in the areas of impact assessment, evaluation and learning in development. The Centre's work is presently focused on:

(1) Exploring a broader range of evaluation designs and methods, and approaches to causal inference.
(2) Designing appropriate ways to assess the impact of complex interventions in challenging contexts.
(3) Better understanding the political dynamics and other factors in the evaluation process, including the use of evaluation evidence.

This CDI Practice Paper was written by **Adinda Van Hemelrijck** and **Irene Guijt**.

ISSN: 2053-0536
AG Level 2 Output ID: 323

Institute of Development Studies, Brighton BN1 9RE, UK
T +44 (0) 1273 915637 **F** +44 (0) 1273 621202 **E** ids@ids.ac.uk **W** www.ids.ac.uk